

From the Public Web to Machine Knowledge: Common Crawl and the Information Infrastructure of Artificial Intelligence

Douglas C. Youvan

doug@youvan.com

www.youvan.ai

July 29, 2026

Artificial intelligence is commonly described through models, algorithms, computing power, and human-like capabilities. Less attention is given to the information infrastructure from which machine intelligence is constructed. Common Crawl occupies a central position within that infrastructure. The nonprofit organization repeatedly collects large portions of the publicly accessible web and distributes the resulting archive for research and computational analysis. Its repository does not constitute a verified account of reality, a finished artificial-intelligence training set, or a neutral sample of human knowledge. It is a vast intermediate layer through which public expression becomes machine-readable data. Developers transform portions of this archive through extraction, filtering, deduplication, classification, weighting, and tokenization before using them to train language models. Each step introduces consequential judgments about quality, relevance, authority, permission, and representation. Common Crawl therefore reveals that artificial intelligence does not simply learn from “the internet.” It learns from selected reconstructions of the internet. This paper examines Common Crawl as public infrastructure, analyzes its role in the artificial-intelligence data supply chain, and considers its implications for epistemic authority, cultural memory, information warfare, synthetic content, copyright, scientific reproducibility, and the future relationship between human publication and machine knowledge.

Keywords: Common Crawl, artificial intelligence, large language models, web archives, training data, machine knowledge, information infrastructure, data curation, synthetic media, epistemology.

Google Drive: <https://drive.google.com/drive/folders/15Kd8bUjm7zVaKdVlwbPhqjiV68aUZQvc>

A collaboration with OpenAI GPT-5.6 Thinking. CC4.0.

<https://orcid.org/0009-0004-4206-4253>

1. Introduction

Public discussion of artificial intelligence often begins too late in the causal chain. Attention is directed toward the finished model: its architecture, number of parameters, computational requirements, benchmark performance, or apparent capacity for reasoning. Yet a language model cannot learn from the world directly. It learns from representations of the world assembled into datasets. Before a model can generate an answer, summarize a controversy, imitate a writing style, or infer a relationship, some collection process must first determine what textual material is available for it to encounter.

Common Crawl is one of the most important systems operating at this upstream level. It is a nonprofit organization that maintains a freely accessible repository of web-crawl data. Its archive contains hundreds of billions of captured pages accumulated across many years, and its recurring crawl cycles add billions of pages at a time. The organization describes its purpose as making large-scale extraction, transformation, and analysis of open-web data available beyond the small number of corporations capable of operating global web crawlers independently.

Common Crawl should not be understood merely as a large collection of saved websites. It is better understood as an infrastructural conversion system. It takes material written for human readers and converts it into forms suitable for bulk computation. A newspaper article, government announcement, scientific paper, commercial advertisement, personal blog entry, political statement, religious essay, discussion-board comment, and automatically generated spam page may all become records within the same underlying archive. Their original purposes differ, but at the level of large-scale data processing they become retrievable computational objects.

This conversion is consequential because artificial-intelligence developers can use Common Crawl as raw material for constructing model-training corpora. The influential Colossal Clean Crawled Corpus, commonly known as C4, was produced by applying a series of cleaning and filtering procedures to a Common Crawl snapshot and was used in the development of Google's Text-to-Text Transfer Transformer, or T5. More recently, the FineWeb project processed 96 Common Crawl snapshots into a corpus containing approximately 15 trillion tokens, demonstrating how an open web archive can be transformed into a modern large-scale pretraining dataset.

These examples do not mean that Common Crawl is itself an artificial-intelligence training set. Nor do they mean that every page captured by Common Crawl has been used to train a particular model. Common Crawl is an upstream source. Between the archived web and the trained model lies an elaborate sequence of technical and institutional decisions. Developers choose which crawl releases to download, which languages to retain, which documents to

remove, how to identify duplication, what constitutes spam, which sources appear educational, which forms of expression are undesirable, and how often particular documents should be represented during training.

The model therefore does not learn “the web” in any direct or comprehensive sense. It learns from a reconstructed web produced by crawling, extraction, filtering, and optimization. The distinction is fundamental. The public web is already a selective representation of human life. A web-derived training corpus is a further selection from that representation. A trained model is then a statistical transformation of that selected corpus. What appears to the user as a direct answer may thus be the downstream result of many layers of exclusion, compression, repetition, and institutional judgment.

Common Crawl makes this hidden genealogy visible. It allows artificial intelligence to be studied not only as a computational artifact but also as an epistemic supply chain. At one end of that chain are people and institutions publishing information. At the other end are models producing apparently synthesized knowledge. Between them lie crawlers, archives, file formats, indexes, filters, classifiers, legal policies, engineering choices, and economic incentives.

The central argument of this paper is that Common Crawl has helped transform the public web from a collection of human-readable publications into a reusable substrate for machine knowledge. This transformation expands access to large-scale data and enables research that would otherwise remain concentrated within major technology companies. At the same time, it allows the inequalities, distortions, commercial incentives, propaganda operations, linguistic imbalances, and unresolved ownership claims of the web to enter artificial-intelligence systems at unprecedented scale.

Understanding Common Crawl is therefore essential to understanding what contemporary artificial intelligence knows, what it fails to know, and why its apparent knowledge cannot be separated from the infrastructure through which information was made available.

2. Common Crawl as Public Data Infrastructure

Common Crawl was founded on a simple but powerful premise: large-scale web data should not be controlled exclusively by the largest search and technology companies. Crawling the internet is technically demanding. It requires discovering URLs, scheduling requests across millions of servers, respecting access restrictions, managing duplicate content, storing enormous quantities of data, maintaining indexes, and distributing the resulting material. An organization that must construct this entire system before conducting research faces a substantial barrier to entry.

By performing much of this collection centrally and releasing the results openly, Common Crawl reduces that barrier. Researchers can investigate language use, web structure, cybersecurity, misinformation, historical change, domain relationships, and machine learning without first operating a global crawler. The archive therefore functions as a shared observational instrument for the web.

The analogy to a scientific observatory is useful. An astronomical observatory gathers signals that many investigators can later analyze for different purposes. Common Crawl similarly gathers digital observations from the public web. The archive does not dictate a single interpretation of those observations. The same crawl can support linguistic research, economic analysis, network science, digital history, cybersecurity investigations, or artificial-intelligence training.

The analogy, however, has limits. A telescope observes physical signals that exist independently of the observer's publication strategy. A web crawler surveys an environment intentionally created by human beings, corporations, governments, automated systems, and strategic actors. The objects being observed may be designed specifically to attract indexing, influence ranking, sell products, persuade readers, conceal sponsorship, or manipulate later computational systems. Common Crawl is therefore not observing a passive natural landscape. It is sampling an active and adversarial information environment.

Its crawler, CCBot, identifies itself to web servers, obeys robots.txt directives, and limits the rate at which it requests pages from individual servers. Website operators may block the crawler or request a slower crawl rate through standard instructions. Common Crawl consequently records only material that is both discoverable and technically accessible under its collection procedures.

This immediately produces a distinction between the public web and the crawled web. A page may be visible to a human reader but inaccessible to CCBot because it requires authentication, payment, interactive scripts, geographic permission, or specialized browser behavior. Other pages may remain uncrawled because few external links point toward them. Conversely, highly connected domains, frequently updated sites, and industrial-scale publishers may be represented repeatedly.

Common Crawl is therefore not a census of the internet. It is a recurring sample shaped by discoverability, technical compatibility, access rules, link structure, scheduling constraints, and crawl priorities. Its extraordinary size should not be mistaken for completeness. Billions of records may still embody systematic absence.

The archive is distributed through several complementary forms. Web ARChive, or WARC, files preserve raw responses received from servers. WAT files contain metadata and structural

information, including links. WET files contain extracted plain text that is especially useful for language processing. Common Crawl also publishes indexes that allow users to determine whether particular URLs were captured and where the corresponding records are stored.

These formats illustrate the gradual abstraction through which a webpage becomes data. The original page may contain typography, images, menus, sidebars, authorship information, advertising, citations, and visual relationships among elements. Plain-text extraction removes much of that context. What remains may be easier for a machine to process, but it may also be detached from signals that helped a human reader judge meaning and credibility.

A headline may become separated from its publication date. A quotation may become separated from its speaker. A correction notice may lose its visual prominence. An advertisement may resemble an editorial paragraph. A satirical claim may survive while the design cues signaling satire disappear. A list of references may become indistinguishable from the article that cites them.

The transformation from webpage to corpus is thus not merely a change in storage format. It is an epistemic transformation. Information becomes more computationally accessible while potentially becoming less contextually complete.

Common Crawl's public character makes this transformation broadly available. That openness is one of its greatest contributions. It permits independent examination of web-scale data and supports reproducible research into datasets that might otherwise remain proprietary. But openness also means that the archive may be used for purposes far beyond those imagined by the original authors of the captured pages.

The same infrastructure that enables academic research can support commercial model development. The same archive that helps investigate propaganda can supply material generated by propagandists. The same dataset that democratizes artificial-intelligence research can reproduce inequalities embedded in online publication. Common Crawl is therefore best understood neither as inherently beneficial nor inherently harmful. It is enabling infrastructure whose consequences depend upon the social, technical, and political systems built upon it.

3. The Artificial-Intelligence Data Supply Chain

The phrase “trained on the internet” compresses a complex process into a misleadingly simple statement. Artificial-intelligence models are not generally connected to an undifferentiated internet during training. Developers construct corpora through a sequence of transformations, each of which determines what the model may learn and how strongly particular patterns are represented.

A simplified sequence begins with public publication. Individuals and institutions place documents online. Crawlers discover some of those documents and preserve copies. Extraction software converts the collected pages into text and metadata. Filters remove material judged unusable or undesirable. Deduplication systems reduce repeated content. Language classifiers divide documents by language. Quality classifiers rank or exclude documents according to selected criteria. Tokenizers convert the retained text into units that a model can process. Training software then repeatedly exposes the model to these units while adjusting its internal parameters.

This process may be summarized conceptually as:

Publication → discovery → crawling → archiving → extraction → filtering → deduplication → weighting → tokenization → training → model behavior

Each arrow represents a possible source of distortion. Material may never be published. Published material may not be discovered. Discovered material may be blocked. Crawled material may be incorrectly extracted. Extracted text may be removed by a filter. Retained text may receive little statistical weight. Learned patterns may not survive later model alignment or may not be reliably recovered during generation.

Consequently, finding a webpage in Common Crawl proves only that the page entered an early stage of potential availability. It does not prove that the page was used to train a particular model, that it survived processing, that its claims were accepted, or that it materially affected any answer.

Yet entry into the upstream supply chain remains significant. Once a page has been incorporated into an openly available web archive, it becomes easier for unknown future actors to process it. Its original website may disappear, but the archived record may remain available in a crawl release or in a derived dataset. Publication thereby acquires a temporal reach beyond the immediate audience of the page.

This is one of the central transformations introduced by web-scale artificial intelligence. Public writing is no longer addressed only to present human readers. It may also become latent instructional material for future machines.

4. The Web Is Not the World

The scale of Common Crawl can create an illusion of comprehensiveness. When a dataset contains billions of pages, it is tempting to imagine that it approximates the totality of human knowledge. Yet size and representativeness are different properties. A vast collection may still omit entire populations, languages, institutions, traditions, and forms of experience.

The web is not a neutral mirror of civilization. It is a historically specific environment produced by unequal access to electricity, telecommunications, literacy, wealth, institutional power, and technical expertise. Some communities generate immense quantities of searchable text. Others preserve knowledge orally, locally, privately, ritually, or in formats that are difficult to crawl. Some governments publish extensive digital archives. Others restrict public access, censor communication, or lack the resources to digitize records. Some languages support large commercial publishing ecosystems. Others have only limited digital representation.

Artificial intelligence trained on web-derived corpora therefore encounters humanity through a highly uneven documentary surface. What is abundant online may become statistically prominent in a model even when it is socially marginal, commercially manufactured, or factually unreliable. What is scarce online may become weakly represented even when it is historically important, locally authoritative, or deeply meaningful to the people who possess it.

This imbalance produces a fundamental epistemic distortion. Machine learning systems infer patterns from frequency, association, and context. They do not naturally distinguish between the frequency of a statement and the importance of the reality to which it refers. A subject supported by millions of pages may produce a strong and flexible representation. A subject preserved in a small number of specialized documents may produce a fragile one. The model may consequently speak fluently about heavily published matters while treating poorly digitized knowledge as obscure, uncertain, or nonexistent.

The problem is not solved simply by increasing the size of the crawl. Larger crawls may reproduce the same structural inequalities at greater scale. If the underlying web favors well-funded institutions, dominant languages, wealthy populations, and professional publishers, then collecting more of the web may intensify those advantages rather than correct them.

This distinction can be expressed as a general principle:

More data from an unrepresentative environment does not necessarily produce a more representative model.

The principle applies beyond geography and language. The web also overrepresents people and organizations motivated to publish. Those with no incentive to produce public text remain comparatively invisible. Commercial actors publish to attract customers. Political organizations publish to recruit supporters. governments publish to establish legitimacy. Advocacy groups publish to influence debate. Search-engine optimization firms publish to capture traffic. Automated systems publish because the marginal cost of producing another page is almost zero.

The resulting corpus is shaped not only by what people know, but also by what they want others to notice.

A model trained on this environment may inherit a systematic confusion between knowledge and publication intensity. It may encounter thousands of pages repeating a commercially useful claim and only a handful of pages documenting a less profitable correction. It may see extensive institutional language describing an official policy while receiving little testimony from the people affected by it. It may learn the vocabulary of administrative systems more deeply than the lived experiences those systems govern.

The public web also favors the present over the past. Recent documents are more likely to remain technically accessible, formatted for modern browsers, and linked from active websites. Historical materials may exist as scanned images, poorly extracted PDFs, damaged archives, or uncatalogued documents. Even when historical material is available, the model may encounter contemporary interpretations of the past more frequently than primary evidence from the period itself.

This creates another layer of mediation. A model may not learn an event principally from records produced when the event occurred. It may learn it from later summaries, textbook language, political reinterpretations, institutional commemorations, and popular retellings. The past reaches the model through the publication priorities of the present.

Common Crawl does not create these inequalities, but it operationalizes them. It transforms the uneven visibility of the public web into a form that can be measured, filtered, and used at scale. In doing so, it converts social asymmetries into computational asymmetries.

The result is not a machine that knows humanity in proportion to human significance. It is a machine that encounters humanity in proportion to documentary availability, crawlability, repetition, and retention.

5. Filtering as a Hidden Constitution

Raw web data cannot be used uncritically. It contains duplicated pages, navigation menus, advertisements, broken text, malicious code, pornography, fabricated information, automated spam, machine-translated fragments, personal data, and vast quantities of low-value material. Any serious artificial-intelligence training pipeline must filter the archive.

Filtering is often described as data cleaning. The phrase suggests a technical process in which contamination is removed while the underlying information remains intact. In reality, filtering is closer to constitutional design. It establishes the rules governing which forms of language may enter the model's world.

A filter must distinguish desirable from undesirable text. It may remove documents containing certain words, pages with particular structural features, text that resembles spam, content

judged sexually explicit, documents written in unsupported languages, or pages whose prose is classified as low quality. Other systems may assign scores and retain only documents above a threshold.

Each of these decisions expresses assumptions about value.

A classifier trained to recognize “high-quality” text may favor standardized grammar, conventional punctuation, institutional vocabulary, and familiar essay structure. It may penalize dialect, fragmented testimony, experimental literature, informal communication, translated prose, or writing produced outside professional institutions. A system intended to remove incoherent text may also remove poetry, code-switching, ritual language, or highly technical notation. A profanity filter may exclude abusive material while also suppressing documents that analyze abuse, discrimination, war, sexuality, or criminal conduct.

The difficulty is structural. The characteristics that distinguish valuable unconventional material from noise are often contextual. A rare technical term may resemble corruption. A witness account may resemble unedited speech. A historical document may contain offensive language because it records the prejudices of its period. A medical discussion may contain terminology that resembles explicit content. A political dissident’s writing may resemble propaganda to a classifier trained on mainstream institutional sources.

Filtering therefore cannot be reduced to the removal of objectively defective data. It is an act of boundary formation.

The training corpus becomes a constructed intellectual environment. Its designers determine which languages are sufficiently represented, which sources are trusted, which subjects require exclusion, which repetitions count as duplication, and which documents appear useful for reasoning. These choices need not be ideologically motivated to have ideological consequences. A seemingly neutral preference for polished prose can privilege institutions that employ editors. A preference for highly cited sources can reinforce established authorities. A preference for widely linked pages can magnify already dominant voices.

The model’s apparent neutrality may conceal these upstream selections. Users encounter a conversational system that presents answers in a uniform voice. They do not ordinarily see which documents were removed, which survived, which received greater weight, or which definitions of quality governed the process. The resulting model can appear universal precisely because the exclusions that produced it are invisible.

This hidden constitution operates at several levels.

First, collection determines which parts of the web are available.

Second, extraction determines which features of the original page survive.

Third, filtering determines which documents remain.

Fourth, weighting determines how frequently the model encounters the retained documents.

Fifth, alignment determines which learned tendencies may be expressed in interaction.

The final answer is therefore shaped by multiple constitutions layered upon one another. The crawler governs access. The extractor governs form. The filter governs admission. The training procedure governs statistical influence. The alignment system governs expression.

No single stage fully determines the model, but each constrains the next.

This layered structure complicates claims of bias. A problematic answer may not arise from one identifiable document. It may emerge from the relative abundance of thousands of documents, the exclusion of countervailing material, the weighting of certain sources, or the interaction between pretraining and later alignment. Conversely, the removal of an individual page may have no measurable effect if the same language appears throughout the corpus.

The unit of influence is therefore often not the isolated document but the patterned environment.

This is why transparency about training data cannot be reduced to publishing a list of URLs. A useful account would also describe collection dates, extraction methods, language detection, quality scoring, deduplication, exclusion categories, source weighting, synthetic-text detection, and later alignment procedures. Without this information, the apparent genealogy of the model remains incomplete.

Common Crawl makes large-scale research possible because its archive is open. Yet the datasets derived from it may be far less transparent. The original crawl can be inspected, but the exact path from that crawl to a commercial model may remain proprietary. Public information enters an open archive, passes through a private curation process, and emerges as a model whose internal representations cannot be traced back cleanly to individual sources.

The public web thereby becomes raw material for privately constituted machine knowledge.

6. Repetition, Coordination, and the Manufacture of Apparent Consensus

One of the most serious implications of web-derived artificial intelligence concerns repetition. Machine-learning systems are sensitive to recurring patterns. When a phrase, claim, association, or narrative appears across many documents, the pattern may become easier for the model to learn and reproduce.

Repetition, however, is not equivalent to independent confirmation.

A single press release may be copied by hundreds of news aggregators. A public-relations statement may be reposted across corporate partner sites. A political message may be distributed through a coordinated network of nominally separate domains. A fabricated claim may spread through blogs, social media transcripts, automated summaries, and derivative articles. To a crawler, these may initially appear as many documents. Socially, they may represent one originating source.

Deduplication can remove exact copies, but coordinated messaging rarely requires exact duplication. The same proposition can be paraphrased, translated, summarized, expanded, or embedded within superficially different articles. The surface forms differ while the underlying narrative remains centrally produced.

This creates a problem of false plurality. The corpus may contain what appears computationally to be broad agreement but is historically traceable to a narrow set of actors.

Traditional scholarship addresses this problem through provenance. Researchers ask who produced a claim, what evidence supports it, whether sources are independent, and whether later writers merely repeat earlier assertions. Web-scale training pipelines are not always able to reconstruct these relationships. They may preserve the text while losing the genealogy of the text.

The difference is crucial. Ten independent investigations reaching the same conclusion provide a different kind of evidence from ten websites reproducing one statement. Yet both situations may generate similar textual recurrence.

Artificial intelligence therefore needs more than content recognition. It requires source-relationship recognition. A trustworthy system would ideally distinguish among independent observation, syndicated publication, common ownership, coordinated campaigns, quotation, citation, criticism, and repetition without verification.

Common Crawl contains some of the information needed for such analysis. Link structures, domains, timestamps, and raw page records can help investigators reconstruct relationships among sources. But these signals are not automatically preserved in the simplified text corpora commonly used for training. Plain-text extraction may retain the claim while discarding the metadata required to understand how the claim propagated.

This creates an epistemic inversion. The information most convenient for statistical learning may be the information least sufficient for judging credibility.

The problem becomes more acute when organizations intentionally design content for machine discovery. Search-engine optimization already rewards pages that use predictable headings, repeated keywords, internal links, and frequently updated text. As AI answer systems become

more influential, publishers may adapt their strategies to influence not only search ranking but also machine retrieval and future model training.

This emerging practice might be called **machine-directed publication**.

Machine-directed publication treats automated systems as the primary audience. The page may be structured to maximize crawlability, extraction quality, semantic clarity, or likelihood of citation by an answer engine. Declarative sentences may be repeated across a network.

Questions may be phrased in forms that resemble user prompts. Claims may be surrounded by the language of explanation and authority. Multiple websites may present the same narrative through different branding.

The purpose is not necessarily to persuade a human reader who carefully examines the evidence. It may be to increase the probability that a machine encounters, retrieves, summarizes, or learns the desired association.

This does not guarantee success. Modern training systems use deduplication, quality filters, source ranking, and other defenses. Retrieval systems may prioritize established domains. Model behavior is influenced by enormous corpora, and the effect of any one campaign may be negligible.

Nevertheless, the incentive exists. As machine-generated answers become gateways to information, actors will seek to shape the informational environments from which those answers arise.

The resulting contest extends beyond conventional propaganda. Traditional propaganda targets human perception. Machine-directed influence targets the infrastructure that mediates human perception. Its object is not only the reader but also the crawler, index, training pipeline, embedding system, retrieval engine, and model.

This creates a new strategic chain:

Publish the narrative → ensure crawlability → multiply its presence → enter indexes and corpora → increase retrieval or training exposure → influence machine-mediated answers

Every step is probabilistic. No single archived page ensures a changed response. Yet coordinated actors do not require certainty. They require only the possibility that sufficient scale, repetition, and technical optimization will shift the informational field.

The risk is greatest when users treat fluent synthesis as independent judgment. A model may summarize a repeated narrative without revealing that many of its apparent sources share a common origin. It may sound as though it has surveyed a broad debate when it has actually inherited a concentrated publication campaign.

The defense against this problem cannot be limited to removing false statements. Some influence campaigns operate through selective emphasis rather than outright fabrication. They publish true facts while omitting context, repeat favorable interpretations, redefine disputed terms, or flood the web with one side of an argument. The resulting corpus may contain few individually false sentences while still presenting a systematically distorted world.

The central issue is therefore not merely truth at the sentence level. It is the architecture of informational visibility.

Common Crawl reveals this architecture because it stands at the point where visibility becomes data. It shows how the public web can function simultaneously as a record of human communication and as a contested training environment for artificial intelligence.

The deeper lesson is that machine knowledge is vulnerable wherever publication volume can substitute for evidentiary independence. A model that cannot distinguish recurrence from corroboration may confuse organized presence with social consensus and coordinated messaging with distributed knowledge.

7. The Synthetic-Web Feedback Loop

The public web is undergoing a transformation that may be as consequential as its original expansion. For most of its history, the web consisted predominantly of material written, photographed, recorded, or deliberately assembled by human beings. Automated systems certainly generated advertisements, product listings, financial records, weather reports, and spam, but the principal linguistic substrate remained human expression. Generative artificial intelligence is altering that balance.

Language models can now produce articles, summaries, reviews, comments, product descriptions, political messages, instructional pages, and entire websites at negligible marginal cost. A single person or organization can publish more text in one day than a conventional editorial staff might previously have produced in months. Much of this material is not labeled as synthetic. Once published, it becomes externally similar to human-authored web content and may later be encountered by crawlers such as CCBot.

This produces a recursive information cycle:

Human publication → web crawl → AI training → AI-generated publication → later web crawl → later AI training

The cycle does not mean that every AI-generated page will enter a future model. Crawlers may miss it, filters may reject it, deduplication may remove it, or developers may deliberately

exclude synthetic text. Nevertheless, the boundary between source material and model output is becoming increasingly difficult to preserve. The web from which future systems learn will contain growing quantities of language produced by earlier systems.

This creates a problem deeper than ordinary duplication. A copied human document retains its connection to an original human source, even when that connection is obscured. Recursively generated text is a statistical derivative of prior corpora. It may reproduce ideas from thousands of sources without preserving their identities, combine true and false statements, regularize unusual language into familiar forms, and present uncertain associations as coherent exposition.

When this material reenters the training environment, the model is no longer learning only from human observations and interpretations. It is learning partly from previous models' transformations of those observations and interpretations.

Research on recursive training has shown that indiscriminate reliance on model-generated data can produce what has been called **model collapse**. In this process, successive generations of models progressively lose features of the original data distribution. Rare events and unusual expressions are especially vulnerable because generative systems tend to reproduce high-probability patterns more often than low-probability ones. When later systems learn from those outputs, the diminished tails of the distribution may shrink further.

The most visible endpoint of collapse may be degraded or nonsensical output, but complete failure is not the only concern. A system may remain grammatically fluent while becoming epistemically narrower. It may preserve common formulations while losing minority dialects, rare scientific observations, unconventional arguments, regional knowledge, and unusual combinations of ideas.

This is better described as **cultural compression** than simple technical malfunction.

Suppose a human corpus contains many ordinary statements and a small number of rare but meaningful statements. A language model trained on that corpus will learn the ordinary patterns more reliably. When asked to generate new text, it will tend to produce those common patterns. A later corpus built partly from its outputs will consequently contain an even greater proportion of ordinary formulations. If this process is repeated, the informational environment may become increasingly centered on what earlier systems already regarded as probable.

The result is a feedback loop in which statistical normality reproduces itself.

Such a loop could make future models appear smoother, more coherent, and more standardized while reducing their contact with the full diversity of human expression. Fluency may survive

even as informational breadth declines. The system may speak more confidently because its world has become more homogeneous.

Model collapse is not inevitable whenever synthetic data is used. Artificially generated examples can be valuable when they are carefully designed, verified, combined with authentic data, and targeted toward known deficiencies. Later research has shown that curation, weighting, pluralistic objectives, and continued access to real data can mitigate at least some recursive-training failures.

The crucial distinction is between **synthetic data as a controlled instrument** and **synthetic data as uncontrolled environmental contamination**.

In a controlled setting, developers know which examples were generated, which model produced them, what purpose they serve, and how their accuracy was evaluated. On the open web, those properties are often unknown. A page may have been written by a human, generated entirely by a model, lightly edited by a human, translated by a model, expanded from a human outline, or assembled from multiple automated systems. Its visible text may not disclose its provenance.

Common Crawl therefore faces a changing observational problem. It was designed to collect the public web, but the ontological status of the public web is becoming less stable. A crawl no longer captures only human communication and conventional automation. It captures an evolving hybrid environment in which models produce part of the material from which later models may learn.

This hybridization complicates the interpretation of scale. A future crawl may contain more pages and more tokens than any previous crawl while containing less independent human information. An increase in textual volume may conceal a decrease in informational originality.

The distinction between volume and novelty becomes essential. Ten million synthetic paraphrases of existing material do not provide the same epistemic value as ten million independently produced observations. Yet ordinary corpus statistics may count both as additional text.

Synthetic proliferation can also alter the apparent distribution of opinion. Suppose a political, commercial, or ideological organization generates thousands of articles expressing one interpretation of an event. Even when the pages differ at the sentence level, their arguments may derive from the same instructions and source packet. A crawler records a large textual population. The social reality, however, may be one centrally directed publication operation.

The synthetic web therefore intensifies the false-plurality problem. It becomes possible to manufacture not only repeated claims but also artificial variation around those claims. Each

page can have a different title, structure, vocabulary, and apparent authorial voice while preserving the same underlying narrative.

Conventional deduplication may fail because the documents are not exact copies. Semantic clustering may identify some relationships, but sufficiently varied generation can make coordinated material resemble independent production. The cost of creating artificial consensus falls dramatically.

A further risk is error sedimentation. A model invents a date, citation, quotation, or causal explanation. The output is published online. Other pages copy or paraphrase it. A later search system retrieves those pages as apparent confirmation. A future model then encounters the claim in several sources and learns it as part of the web's recurring pattern.

The error has moved through several stages:

Generation → publication → repetition → retrieval → apparent corroboration → possible retraining

At the end of the process, the false claim may appear older, broader, and more established than it actually is. Its synthetic origin becomes difficult to reconstruct.

This is not entirely new. Human societies have always repeated rumors, legends, propaganda, and mistaken citations. The difference is one of scale, speed, variation, and automation. Generative systems can produce large families of mutually reinforcing documents without requiring a corresponding increase in human belief, investigation, or experience.

Common Crawl is not responsible for this transformation, but it may preserve its products. Its archives may eventually become valuable precisely because they record earlier periods in which the proportion of human-created material was greater. Historical crawl releases could function as informational reference layers against which later synthetic proliferation is measured.

The archive may therefore acquire a second role. It is not only a source for training future artificial intelligence. It may also become evidence of what the web looked like before artificial intelligence substantially rewrote it.

8. Provenance and the Disappearance of Informational Genealogy

Knowledge depends not only on what is said but also on where a statement came from, who produced it, what evidence supported it, and how it changed during transmission. These relationships constitute provenance.

In scholarship, provenance appears through citations, authorship, publication history, archival custody, experimental methods, and chains of quotation. In journalism, it appears through named sources, documentary evidence, editorial responsibility, and corrections. In law, it appears through testimony, exhibits, precedents, and chains of custody. In science, it appears through methods, data, laboratory records, and reproducible procedures.

Provenance allows a reader to distinguish between an observation and a rumor, an original source and a later summary, an independent confirmation and a copied statement.

Web-scale language processing often weakens these distinctions.

When a webpage is converted into extracted text, many of its genealogical signals may disappear. The plain text may preserve the body of an article while omitting visual indicators of authorship, sponsorship, correction status, or quotation hierarchy. Links may be removed or stored separately. Publication dates may become ambiguous. Captions may detach from images. A document quoting a false claim in order to refute it may be reduced to text containing both the claim and the refutation without a reliable representation of their relationship.

Training then compresses the retained material into model parameters. The model does not ordinarily preserve each statement as a discrete record attached to a complete source history. It learns distributed statistical relationships across many documents. This enables broad generalization, but it makes exact attribution difficult.

A model may know a proposition without being able to identify where it learned it. It may generate an accurate statement from a mixture of sources yet cite an unrelated page. It may reproduce the wording of a source without recognizing that the wording is distinctive. It may combine a claim from one document, a date from another, and an interpretation from a third into a sentence that appeared in none of them.

The result is **knowledge without visible genealogy**.

This feature separates parametric machine knowledge from traditional documentary knowledge. A library stores identifiable documents. A search engine returns identifiable pages. A conventional database stores records with fields and origins. A language model stores learned relationships in a distributed form that does not naturally reveal which source caused which output.

The problem is not merely one of intellectual credit. Provenance is necessary for epistemic evaluation. A claim derived from a government press release should be assessed differently from the same claim derived from an independent investigation. A medical assertion from a peer-reviewed study should be distinguished from an advertisement repeating the study's

conclusion. A historical quotation should be traceable to a primary document rather than to a chain of unsourced websites.

Without provenance, all three may enter the model as recurring linguistic associations.

Researchers have proposed claim-provenance graphs and source-aware training methods that attempt to preserve relationships among claims, documents, transformations, and responsible parties. These approaches recognize that natural-language provenance is frequently uncertain and must itself be inferred from textual and network evidence.

Common Crawl contains raw materials that could support richer provenance analysis. WARC records preserve server responses. WAT records retain metadata and links. Crawl timestamps indicate when pages were collected. Domain relationships, hyperlinks, identical passages, and publication sequences can help reconstruct how a claim traveled.

Yet these properties may be lost when developers reduce the archive to disconnected text documents. The most efficient format for training may not be the most adequate format for preserving epistemic history.

This tradeoff can be stated directly:

Text extraction increases semantic usability while decreasing documentary completeness.

A corpus designed primarily for next-token prediction may retain enough language to teach grammar, concepts, and associations while discarding information needed to judge independence, authority, and intent. The model learns what statements tend to occur but not necessarily why they occurred or how they relate to their sources.

A stronger AI information architecture would preserve several layers simultaneously.

The first layer would contain the original document as captured.

The second would contain structured metadata describing the domain, author, date, publisher, ownership, and technical circumstances of capture.

The third would represent citations, quotations, syndication, common ownership, and semantic similarity among documents.

The fourth would track transformations performed by extraction, filtering, translation, summarization, and synthetic generation.

The fifth would connect model outputs to the documents or evidence used to support them.

Such a system would treat provenance not as optional decoration but as part of the knowledge itself.

This would be computationally difficult. Authorship metadata can be missing or false. Corporate ownership changes. Websites copy material without attribution. Publication dates are manipulated. Coordinated organizations conceal their relationships. Synthetic text may imitate the style of independent journalism. A complete genealogy of every web claim may be impossible.

Nevertheless, partial provenance is better than none. Uncertainty can be represented explicitly. A system can state that several pages repeat a claim but appear to derive from one press release. It can distinguish a primary document from commentary. It can indicate that a source's ownership is unknown or that an article's apparent date conflicts with its crawl history.

The goal should not be perfect certainty. It should be the preservation of enough informational structure to prevent recurrence from masquerading automatically as corroboration.

Provenance also changes the meaning of correction. In ordinary publishing, a false statement can be corrected at its source. But crawled copies may preserve the earlier version, derived datasets may contain it, and models may already have learned its language. A correction placed on the original page does not necessarily propagate through the downstream supply chain.

The information system therefore requires lineage-aware correction. A corrected claim should ideally be connected to prior versions, derivative pages, archived copies, and downstream datasets. Without that connection, the false statement and its correction simply become two competing textual patterns.

The stronger pattern may prevail statistically, regardless of chronology or evidence.

Common Crawl exposes this problem because it preserves temporal layers of the web. The archive may contain a false original, a corrected version, reproductions of the false version, and discussions of the correction. To a historian, this is a valuable record of informational change. To a training pipeline that treats the records as undifferentiated text, it may be a mixture of contradictory statements.

Time is therefore part of provenance. A model should know not only that a statement exists but also whether it preceded or followed a correction, whether it was current when published, and whether later evidence displaced it.

Without temporal provenance, obsolete information can remain statistically alive.

The long-term challenge is to move from models trained on isolated textual surfaces toward systems capable of reasoning over documentary histories. Such systems would not merely ask, "What do the documents say?" They would ask:

Who first made the claim?

Which later sources copied it?

Which sources independently investigated it?

What evidence was available at each stage?

Was the claim corrected, withdrawn, or superseded?

Who financed or controlled the publication network?

What remains uncertain?

These are not peripheral questions. They are the questions through which information becomes accountable knowledge.

9. Training-Time Influence and Retrieval-Time Influence

Discussions of AI influence often combine two different mechanisms: changing what a model learns during training and changing what a system retrieves while answering a question. The distinction is essential.

Training-time influence occurs when material becomes part of the corpus used to adjust a model's internal parameters. The effect, when it exists, is distributed and often difficult to measure. A document may contribute to patterns learned across billions or trillions of tokens. Its language may be reinforced by similar documents, diluted by contradictory ones, removed by filtering, or rendered inaccessible by later alignment.

Training usually occurs in discrete cycles. Once a model has been trained, publishing a new webpage does not ordinarily rewrite its parameters immediately. The new page can affect a future model only if a developer later collects or acquires it, includes it in a corpus, and retrains or continues training the model.

Retrieval-time influence operates differently. A retrieval-augmented system searches an external index, database, or live web source when responding to a query. Newly published material can potentially affect answers soon after it is indexed. The system may insert retrieved passages into the model's context and ask the model to synthesize an answer from them.

The corresponding chains are different:

Training-time pathway: publication → crawl → corpus selection → model training → latent influence

Retrieval-time pathway: publication → indexing → query match → document retrieval → immediate answer context

The first pathway is slow, diffuse, and difficult to attribute. The second can be rapid, local, and visible.

This distinction changes how claims of AI manipulation should be evaluated. Discovering that a website appears in Common Crawl establishes possible upstream availability. It does not show that a current model was trained on the site. By contrast, observing that an answer engine retrieved and cited the website during a test provides direct evidence of retrieval-time exposure for that query.

Even then, exposure is not equivalent to persuasion. The model may quote the source, criticize it, compare it with others, or use only a minor factual detail. A proper analysis must examine what was retrieved, how prominently it was used, whether independent sources were included, and whether the final answer adopted the source's framing.

Retrieval systems can offer advantages over purely parametric models. They can access recent information, provide citations, update knowledge without retraining the entire model, and allow users to inspect supporting documents. They can also separate general linguistic capability from current factual evidence.

But retrieval creates its own attack surface.

Publishers may optimize pages for likely queries. Coordinated networks may target narrow questions for which authoritative sources are scarce. Search rankings may be manipulated through linking strategies, domain multiplication, or highly specific phrasing. A newly created page may outrank an older primary source if the retrieval system values semantic similarity more than institutional authority or documentary provenance.

Citations do not automatically solve the problem. A citation shows that a source was associated with an answer, but it does not prove that the source supports every claim attributed to it. Recent evaluations of retrieval-grounded systems continue to distinguish between citation presence, citation correctness, claim support, source quality, and completeness of attribution.

A system can therefore appear transparent while remaining epistemically weak. It may provide several links, yet those links may all derive from one original statement. It may cite a reputable page that mentions a subject without supporting the model's specific conclusion. It may attach a citation to a paragraph containing multiple claims even though the source supports only one of them.

The relevant standard is not whether citations are present. It is whether the answer's evidentiary structure is faithful.

A robust retrieval system should perform at least four forms of evaluation.

First, it should determine whether the retrieved document directly supports the claim.

Second, it should assess the quality and role of the source.

Third, it should identify relationships among apparently separate sources.

Fourth, it should represent disagreement and uncertainty rather than forcing all retrieved material into a single seamless narrative.

The difference between training and retrieval also affects removal and correction. A website owner can block future crawling, delete a page, or request removal from a search index. These actions may reduce future retrieval. They cannot reliably erase information already incorporated into a trained model's parameters or remove copies stored in historical archives and derivative datasets.

The web therefore has multiple informational time scales.

Retrieval systems operate near the present. They respond to what is currently indexed.

Training corpora capture selected historical intervals.

Model parameters preserve compressed statistical residues of those corpora.

Archives preserve earlier versions even after the live web changes.

An organization seeking influence may target all four levels. It may publish material for immediate retrieval, seek repeated crawling for future availability, build a durable network of mutually reinforcing pages, and preserve favorable narratives in archives. Conversely, a person seeking correction must confront all four levels: the live page, the index, the archive, and the trained model.

This layered persistence makes informational control difficult. There is no single switch by which a statement enters or leaves artificial intelligence.

The most responsible interpretation of Common Crawl is therefore limited but important. A Common Crawl record demonstrates that a page was captured and made available within a widely used public data repository. It indicates potential accessibility to researchers and developers. It does not demonstrate actual inclusion in a named model, measurable parameter influence, or adoption of the page's claims.

Evidence of retrieval provides a stronger connection to a particular answer, but even retrieval must be analyzed at the level of claim support and source use.

These distinctions protect analysis from both exaggeration and complacency. It is an exaggeration to claim that every crawled page has trained artificial intelligence. It is complacent

to dismiss crawl presence as meaningless merely because direct influence cannot be proved. Entry into a durable, reusable data infrastructure creates future possibility, even when the downstream path remains unknown.

Common Crawl matters because it lowers the distance between publication and computational reuse. Retrieval systems lower that distance further. Together, they transform the web into both a historical training reservoir and a near-real-time evidentiary environment.

The emerging information contest will therefore occur along two fronts. One concerns what future models internalize. The other concerns what present systems retrieve. Understanding artificial intelligence requires keeping both mechanisms visible.

10. Copyright, Consent, and the Meaning of Public Access

Common Crawl operates within a conceptual ambiguity that has become central to the politics of artificial intelligence: the difference between information being publicly accessible and information being freely available for every subsequent use.

A webpage may be readable without payment or authentication. That fact does not necessarily mean that its author intended the text to be copied into a permanent archive, redistributed in bulk, transformed into a commercial dataset, or used to train a model capable of reproducing its style and substituting for its original market.

Public access is a condition of visibility. It is not automatically a complete theory of permission.

This distinction was less politically prominent when web crawling was associated mainly with search indexing. Search engines copied pages in order to direct users back to those pages. The publisher could plausibly benefit from indexing because the system increased discovery and traffic. The relationship was not always consensual or equal, but it retained a recognizable exchange: the crawler copied enough information to help the reader find the source.

Artificial-intelligence training changes that relationship. A model may learn from the page without later sending the user to it. It may summarize the source, imitate its explanatory structure, answer questions that the source once answered, or combine its contents with other materials into a new response. The user may receive the informational benefit without knowing that the original document existed.

The economic direction of value can therefore reverse. Instead of the index directing attention toward the publisher, the model may absorb the publisher's work and redirect attention away from the original publication.

This does not mean that every use of public web data is legally or ethically equivalent. Search indexing, archival preservation, linguistic research, security analysis, nonprofit scholarship, commercial training, model evaluation, and synthetic generation are distinct activities. A defensible framework must recognize their differences rather than treating all machine processing as one category.

The legal status of these activities varies across jurisdictions and remains contested. Copyright law was not designed around statistical models trained on billions of documents. It contains doctrines addressing copying, transformation, quotation, research, indexing, and fair use, but their application to large-scale training involves unresolved questions.

Several issues recur.

The first concerns the training copy itself. Developers generally create copies of works during dataset construction and training. The legal significance of those copies depends on purpose, transformation, market effect, jurisdiction, and other factors.

The second concerns model outputs. A model may generate text that resembles a source, reproduces a protected passage, imitates a distinctive style, or provides a substitute for the original work. The probability and legal meaning of such output can differ from the legality of training.

The third concerns dataset transparency. An author cannot challenge or evaluate the use of a work when the developer does not disclose whether the work was included.

The fourth concerns removal. Even when a live webpage is deleted or blocked from future crawling, copies may persist in archives, derived datasets, research mirrors, corporate storage, and model parameters.

The fifth concerns collective scale. A single copied page may appear insignificant, yet the commercial value of a model may depend upon the aggregated use of millions of works whose individual creators received no notice or compensation.

The cumulative structure is essential. Artificial intelligence is not trained by negotiating separately with every sentence. Its power comes from aggregation. The law, however, often evaluates claims at the level of identifiable works and plaintiffs. This creates a mismatch between the distributed nature of model construction and the individualized structure of legal remedy.

Consent presents a related but distinct problem.

Website operators can use robots.txt to instruct compliant crawlers not to visit parts of a site. This gives publishers a limited technical mechanism for controlling future collection. Yet such instructions have several limitations.

They apply prospectively rather than retroactively.

They depend on voluntary compliance by crawlers.

They may block beneficial search or archival uses along with AI collection.

They are controlled by the current website operator, who may not be the author or copyright holder.

They cannot remove copies already collected.

They do not communicate nuanced preferences among indexing, archiving, research, and commercial training.

A binary permission system is poorly matched to a world of heterogeneous computational uses.

A publisher may welcome search indexing but reject model training. An author may support nonprofit research but oppose commercial reuse. A newspaper may permit snippets but not full-text extraction. A researcher may want a paper preserved but require attribution in downstream systems. A deceased author's estate may hold different rights from the institution hosting the work.

The future web may therefore require machine-readable licensing that is more expressive than simple crawl permission. Such a system could distinguish among access for search, preservation, academic analysis, commercial training, retrieval, quotation, and synthetic generation.

This would not resolve every dispute. Metadata can be ignored. Ownership can be contested. Rights change over time. Public-interest research may require exceptions. National laws differ. But a richer permissions layer would at least allow creators and institutions to express intentions that current infrastructure cannot represent.

The ethical problem extends beyond copyright. Public webpages can contain personal information, photographs, medical discussions, family histories, contact details, legal records, and statements written without anticipation of machine-scale aggregation. Information dispersed across many pages may be relatively obscure to human readers but become easily searchable when collected into a centralized corpus.

Aggregation changes privacy.

A fact that is technically public but practically difficult to find may acquire a different social meaning when integrated into a dataset or model. Artificial intelligence can connect fragments,

infer relationships, summarize personal histories, and surface material outside its original context. The harm may arise not from the publication of any single fact but from the removal of friction that once limited discovery.

Common Crawl exists at the boundary between openness and aggregation. Its public mission provides substantial benefits. Researchers can study the web without relying entirely on proprietary corporate infrastructure. Historical material can survive the disappearance of live pages. Independent investigators can analyze domain networks, misinformation campaigns, and technical trends.

Yet the same openness distributes captured material beyond the control of original publishers.

The archive therefore presents a genuine conflict between two public goods: broad access to computational research infrastructure and meaningful control over one's work and information.

A simple prohibition on open crawling could entrench the largest corporations. Firms with proprietary crawlers, licensed databases, platform access, and capital-intensive data partnerships would continue collecting at scale. Universities, nonprofit laboratories, journalists, and independent researchers would lose a shared resource.

Conversely, unrestricted crawling and reuse could place the burden of AI development disproportionately on creators, small publishers, and individuals who lack bargaining power.

The policy objective should not be to choose categorically between openness and control. It should be to distribute rights, obligations, and benefits more fairly across the information supply chain.

Possible principles include transparent disclosure of major training sources, enforceable mechanisms for prospective exclusion, clearer treatment of removal requests, protection for public-interest research, documentation of derived datasets, and stronger safeguards against verbatim reproduction and personal-data extraction.

Attribution also deserves renewed attention. Traditional citation systems assume that a reader can be directed to identifiable sources. Parametric models complicate this because their outputs may reflect diffuse learning from many documents. Nevertheless, retrieval systems can cite the evidence used for a particular answer, and developers can document the major corpora used during training.

The impossibility of perfect attribution should not become an excuse for complete opacity.

The deeper issue is that artificial intelligence transforms publication from a communicative act into a potential resource-extraction event. A person may write for readers, but a crawler may treat the writing as data. A community may publish a cultural record, but a developer may treat

it as linguistic material. A scientist may release a paper to support inquiry, but a model may absorb it into a commercial service.

The same document participates in multiple economies of meaning.

A legitimate AI information order must therefore recognize that accessibility, ownership, authorship, cultural significance, and computational usefulness are not interchangeable properties.

11. Democratization and the Reproduction of Power

Common Crawl is often described as a democratizing resource, and this description is justified. Web-scale data collection is expensive. Without a shared archive, only major search engines, governments, telecommunications firms, and the wealthiest technology companies could routinely analyze the public web at global scale.

By publishing crawl data openly, Common Crawl reduces the cost of entry for universities, nonprofit organizations, small companies, independent researchers, and public-interest investigators. It allows groups without proprietary crawling infrastructure to conduct experiments, construct datasets, study internet structure, and test claims about the web.

This democratization is real, but it is incomplete.

Access to raw data does not eliminate the need for storage, bandwidth, technical expertise, and computing power. Common Crawl may be freely available, but processing a large portion of it can still require substantial resources. Downloading hundreds of terabytes, parsing WARC files, deduplicating documents, training classifiers, and constructing trillion-token corpora remain beyond the practical reach of many researchers.

Open availability therefore does not imply equal usability.

The difference resembles the distinction between a public mineral deposit and the capacity to mine it. The resource may be legally available, but institutions with greater capital, infrastructure, and expertise can extract more value from it.

Large technology companies benefit from several additional advantages.

They can combine Common Crawl with proprietary web crawls, user interactions, licensed books, code repositories, private databases, search logs, and platform content.

They can train larger models.

They can run more experiments.

They can employ specialized legal and engineering teams.

They can deploy models to millions of users and collect feedback at scale.

They can update systems continuously and use market revenue to finance the next generation.

An open crawl may lower one barrier while leaving many others intact.

Common Crawl can therefore democratize access to raw material while the broader AI economy remains concentrated. The archive may even indirectly benefit dominant firms by supplying an inexpensive portion of the data required for commercial model development.

This does not make openness futile. It means that data access alone cannot create an equitable artificial-intelligence ecosystem.

Democratization must be evaluated across the entire chain:

Who can publish?

Who gets crawled?

Who can process the crawl?

Who can train a model?

Who can deploy it?

Who receives economic value?

Who bears legal risk?

Who can challenge errors?

Who controls the interface through which users encounter machine knowledge?

At each stage, power may reconcentrate.

The publication layer favors actors with resources to produce and optimize large quantities of content.

The crawling layer favors technically accessible and well-linked domains.

The processing layer favors institutions with computation and expertise.

The model layer favors organizations with capital.

The deployment layer favors platforms with large audiences.

The governance layer favors parties able to influence regulation and litigation.

Common Crawl intervenes primarily at one point: access to crawled data. Its openness is important, but it cannot by itself correct inequalities elsewhere.

The archive nevertheless creates possibilities that proprietary systems do not. Independent researchers can audit portions of the same underlying web data used in model development. They can identify problematic sources, measure duplication, analyze language representation, and study the emergence of synthetic content. They can construct alternative corpora based on different values.

One research group might prioritize underrepresented languages.

Another might preserve historical and scientific documents.

Another might build a corpus excluding commercial content.

Another might emphasize traceable primary sources.

Another might study how propaganda networks reproduce narratives across domains.

The existence of a public archive allows competing interpretations of the web to be tested.

This plurality is epistemically valuable. If all large-scale web data were proprietary, companies could make claims about model quality, safety, and representativeness without enabling independent examination of comparable source material. Public infrastructure makes at least partial replication and criticism possible.

Common Crawl can therefore support a more plural AI ecosystem even when it does not guarantee one.

The distinction between public infrastructure and public governance is important. A resource may be openly available while its policies remain shaped by a small technical organization. Decisions about crawl priorities, exclusion, rate limits, file formats, and removal procedures affect downstream research. These decisions deserve transparency because the archive functions increasingly like a public knowledge utility.

The same is true of derived datasets. A corpus built from Common Crawl may become more influential than the original archive because developers use it repeatedly. The curators of such a corpus effectively decide which parts of the public web become normalized as machine-learning material.

Dataset creation is therefore a form of institutional power.

The dataset builder determines what counts as language, quality, duplication, toxicity, relevance, and educational value. These definitions may be encoded in scripts and classifiers, but they remain normative decisions.

Open-source publication of filtering code can improve accountability. Documentation can reveal inclusion criteria. Data sheets can identify languages, dates, known limitations, and removal procedures. Reproducible pipelines can allow researchers to test alternatives.

Yet complete reproduction may remain expensive. Transparency without practical contestability can become ceremonial. A company may publish enough information to appear open while retaining the computational advantage required to evaluate its claims fully.

Meaningful democratization requires not only access but also institutional capacity. Public universities, libraries, national archives, and nonprofit laboratories may need shared computing infrastructure capable of processing web-scale data. Otherwise, the public owns the raw archive while private firms dominate its interpretation.

Libraries are especially relevant. Their historical mission includes preservation, cataloging, provenance, public access, and stewardship across generations. These functions correspond closely to the emerging needs of machine knowledge. AI datasets require not merely engineering but also archival judgment.

A library-centered approach would ask different questions from a purely commercial pipeline.

What should be preserved?

How should editions be distinguished?

What metadata must remain attached?

How should corrections be represented?

Which communities possess rights over cultural material?

How should access vary by purpose?

What does responsible deletion mean?

How can future researchers reconstruct the history of a claim?

These are longstanding archival problems newly amplified by artificial intelligence.

Common Crawl demonstrates that the web has become a form of civilizational data infrastructure. Its governance should therefore not be left exclusively to market competition or technical convenience. Public institutions have a legitimate role in determining how collective information is collected, preserved, transformed, and used.

The alternative is a paradoxical form of privatization. Human societies publish their knowledge openly. Public crawlers and shared networks make it computationally accessible. Private firms then transform it into proprietary systems that mediate access back to the same societies.

The public supplies the informational substrate; the corporation owns the interface.

Common Crawl can resist this pattern only partially. It keeps a major part of the upstream resource open. Whether downstream intelligence remains open, accountable, and plural depends on additional institutions.

12. Toward a More Accountable AI Information Architecture

The problems associated with Common Crawl do not imply that web-scale data should be abandoned. The public web contains an extraordinary record of scientific knowledge, historical documentation, technical instruction, cultural expression, public debate, and ordinary life. Excluding it entirely would impoverish artificial intelligence and strengthen organizations with access to proprietary alternatives.

The objective should be to improve the architecture through which web data becomes machine knowledge.

Such an architecture would begin by rejecting the fiction that all text is interchangeable. A scientific article, political advertisement, personal blog, government statement, anonymous forum post, corporate press release, and synthetic content farm should not enter a corpus as epistemically equivalent strings.

They may all contain useful language, but they occupy different documentary roles.

A more accountable system would preserve those roles wherever possible.

12.1 Provenance-Preserving Corpora

Training datasets should retain links between extracted text and original documents. This does not require every model parameter to be traced to a single webpage, which may be impossible. It requires the corpus to preserve enough structure for auditing, correction, and source analysis.

Each document record could include:

- crawl date;
- original URL;
- domain and subdomain;
- publication date when available;
- author and publisher metadata;
- language;

- content type;
- ownership information when known;
- links and citations;
- duplication or syndication clusters;
- extraction history;
- filtering score;
- suspected synthetic origin;
- licensing or access signals;
- correction and update status.

These fields would not always be complete or accurate. Their uncertainty should be represented rather than ignored.

A system that says “author unknown” is more honest than one that discards authorship entirely.

12.2 Source-Relationship Modeling

Documents should not be treated as independent merely because they occupy different URLs. Corpus builders should identify likely syndication, common ownership, shared press releases, quotation chains, translated copies, and coordinated semantic clusters.

This would allow models and retrieval systems to estimate not only how often a claim appears but how many independent source lineages support it.

A useful distinction would be:

Document count $\neq$ source count $\neq$ evidence count

A thousand pages may contain one source and no independent evidence. Three documents may contain three separate investigations. Numerical repetition alone cannot distinguish them.

Source-relationship modeling would not eliminate manipulation, but it would raise the cost of manufacturing apparent consensus.

12.3 Temporal Knowledge

Training corpora should represent when claims were current, corrected, disputed, or superseded. A statement can be accurate at one date and false at another. Policies change. officeholders change. scientific conclusions are revised. companies fail. conflicts begin and end.

A model trained without temporal structure may blend incompatible periods.

Temporal metadata would allow systems to distinguish:

- original publication from later revision;
- rumor from subsequent correction;
- preliminary finding from later replication failure;
- historical officeholder from current officeholder;
- old product specification from current specification;
- contemporaneous testimony from retrospective interpretation.

Artificial intelligence should not merely know statements. It should know their historical interval of validity.

12.4 Synthetic-Content Identification

Future corpora will require increasingly sophisticated methods for identifying machine-generated and machine-transformed text. Binary classification may be impossible because authorship is becoming hybrid. A document may contain human writing, automated translation, model-generated expansion, and editorial revision.

The relevant objective is not always exclusion. Synthetic material can be useful. The objective is provenance and control.

A corpus should distinguish, when possible, among:

- wholly human-authored text;
- human text with automated correction;
- translated text;
- AI-assisted drafting;
- model-generated text with human editing;
- fully synthetic text;
- unknown origin.

The model can then be trained with deliberate weighting rather than accidental recursion.

Preservation of human-origin corpora may become strategically important. Earlier Common Crawl releases, digitized books, recorded oral histories, archival newspapers, and verified

human datasets could serve as reference reservoirs protecting future models from excessive synthetic feedback.

12.5 Transparent Filtering

Dataset curators should disclose how documents were removed or ranked. Transparency should include more than a list of prohibited categories. It should describe classifiers, thresholds, benchmark sources, language coverage, deduplication methods, and known error patterns.

Researchers should be able to ask:

Which dialects were disproportionately removed?

Which domains dominate the retained corpus?

How much text originates from commercial content farms?

What fraction appears synthetic?

Which languages lost the greatest share during quality filtering?

How much repeated content survives semantic deduplication?

Which institutions define the high-quality reference set?

These questions turn filtering from an invisible technical step into an auditable intellectual process.

12.6 Separation of Linguistic and Evidentiary Training

Language models need broad exposure to writing in order to learn grammar, style, concepts, and reasoning patterns. Factual answering, however, requires reliable evidence.

A more mature architecture may distinguish these functions.

Broad web corpora can support linguistic competence.

Curated source collections can support factual grounding.

Retrieval systems can provide current evidence.

Provenance graphs can support source evaluation.

Formal databases can support precise records.

Human experts can adjudicate disputed or high-stakes claims.

The model should not be expected to derive all epistemic authority from the same undifferentiated corpus used to learn language.

This separation would reduce the temptation to treat every learned association as knowledge.

12.7 Claim-Level Citation

Retrieval systems should cite at the level of individual claims rather than attaching a general source list to a long answer. Each factual assertion should be connected to evidence that directly supports it.

The system should also indicate the role of each source:

- primary evidence;
- official statement;
- independent reporting;
- scholarly analysis;
- commentary;
- allegation;
- disputed claim;
- secondary summary.

Citation quality depends not only on whether a link exists but on whether the relationship between claim and source is accurate.

A system should be able to say that a government claims something without presenting the government's statement as independent verification.

12.8 Correction Propagation

Artificial-intelligence infrastructure needs mechanisms for propagating corrections through archives, indexes, datasets, and retrieval systems.

The original false record should not necessarily be erased, especially when historical preservation matters. Instead, it should be linked visibly to the correction.

A correction-aware system might retain:

- the original claim;
- the date it appeared;
- the correction or retraction;
- the evidence motivating the correction;

- later sources that continued repeating the obsolete version.

This preserves history while preventing the outdated claim from remaining epistemically equivalent to its correction.

12.9 Purpose-Specific Permissions

Web publishers need more expressive ways to communicate permissions. A future standard could distinguish among:

- search indexing;
- archival preservation;
- academic research;
- commercial model training;
- retrieval and quotation;
- full-text redistribution;
- style imitation;
- personal-data processing.

No technical standard can resolve all legal disputes, but a granular permissions layer would improve communication and accountability.

The system should also preserve the permissions that applied when the page was collected. A later change in policy should not erase the historical record of what instructions were visible at the time.

12.10 Public Oversight and Shared Capacity

Because web-scale corpora increasingly shape public knowledge systems, their governance should involve more than private developers. Libraries, universities, civil-society organizations, journalists, archivists, technical experts, creators, and representatives of underrepresented language communities should participate in setting standards.

Public investment is also necessary. Independent auditing requires computing resources. It is not enough to release a dataset whose size prevents meaningful examination by all but the organizations that created it.

Shared public infrastructure could support:

- corpus analysis;

- provenance reconstruction;
- multilingual auditing;
- synthetic-content measurement;
- bias assessment;
- historical web preservation;
- removal and correction research;
- public-interest model development.

Common Crawl demonstrates that a shared web-data resource is possible. The next step is shared capacity to interpret and govern that resource.

The ultimate aim should not be a perfectly purified corpus. No such corpus can exist. Human knowledge is contested, incomplete, political, historical, and uneven. Attempts to remove every ambiguity would remove much of what makes knowledge human.

The aim should instead be accountable imperfection.

An accountable system documents its sources, limitations, transformations, uncertainties, and corrections. It does not present statistical fluency as self-validating truth. It allows users and researchers to reconstruct how information entered the system and why one answer was preferred over another.

Common Crawl is valuable precisely because it makes part of the upstream process visible. It shows that machine knowledge begins not inside the model but within an ecology of publication, crawling, storage, curation, and power.

A responsible artificial-intelligence architecture must preserve that ecology rather than hiding it behind the smooth surface of an answer.

13. Conclusion: The Infrastructure Before the Intelligence

Common Crawl appears at first to be a technical utility: a nonprofit crawler, a large archive, and a collection of files made available for computational research. At the scale of contemporary artificial intelligence, however, it becomes something more consequential. It is part of the infrastructure through which the public web is transformed into machine-readable cultural memory.

The distinction between the web and Common Crawl is important. The web consists of materials published for many different reasons by human beings, institutions, governments, businesses, automated systems, and increasingly artificial-intelligence models. Common Crawl observes a portion of that environment according to the practical constraints of discovery, access, crawling, and storage. It is therefore neither a complete copy of the internet nor a representative sample of humanity. It is a recurrent computational survey of what was sufficiently visible, connected, accessible, and technically compatible to be collected.

The distinction between Common Crawl and an AI training dataset is equally important. Common Crawl supplies archived material. Dataset builders then select crawl releases, extract text, classify languages, remove duplication, rank quality, exclude selected categories, and determine how retained documents will be weighted. Modern projects illustrate the magnitude of this transformation. C4 was created by filtering a single Common Crawl snapshot, while FineWeb was constructed from 96 snapshots and contains approximately 15 trillion tokens. These corpora are not merely smaller copies of Common Crawl. They are interpretations of it, produced through thousands of technical and normative decisions.

The resulting artificial-intelligence model is yet another transformation. It does not retain the corpus as a conventional library. It compresses recurring textual relationships into distributed numerical parameters. The final system can generalize beyond individual documents, but it ordinarily cannot reconstruct a complete genealogy for every generated statement. It may know a claim without knowing where the claim originated, whether apparently separate sources were independent, or whether the claim was later corrected.

The complete sequence is therefore not simply web to AI. It is:

Human and institutional publication → web visibility → crawler discovery → archival capture → textual extraction → corpus selection → filtering and weighting → model training → alignment → retrieval → generated answer

Every stage changes the informational object.

A webpage loses some of its visual and documentary context when converted into extracted text. A corpus loses part of the web's diversity when filtered. A model loses explicit documentary boundaries when the corpus is compressed into parameters. An answer loses further context when the model presents a synthesis in a uniform voice.

This sequence explains why artificial intelligence can sound more unified than the world from which it learned. The underlying sources may be contradictory, unequal, strategically produced, and historically layered. The model transforms them into a single linguistic performance. The coherence of the answer may therefore reflect the model's method of presentation rather than the coherence of the evidence.

Common Crawl reveals this problem at its origin. It shows that machine knowledge begins with documentary availability rather than direct perception of reality.

The web is not the world. It is what some portion of the world had the means, incentive, permission, and technical ability to publish. Common Crawl is not the web. It is what its crawler encountered and retained. A training corpus is not Common Crawl. It is what dataset builders selected and transformed. A model is not the corpus. It is a statistical system trained from that transformation. An answer is not the model's complete internal state. It is one generated response conditioned by a particular prompt, alignment policy, and possibly a set of retrieved documents.

At each step, absence can be mistaken for nonexistence.

A culture with little digitized text may appear less intellectually developed than a heavily documented one. A local event that receives little online attention may appear less significant than a trivial event surrounded by commercial publication. A claim repeated across coordinated websites may appear more widely supported than a well-evidenced claim contained in a small number of primary documents. An old error preserved across thousands of derivative pages may compete statistically with a later correction.

Artificial intelligence consequently inherits not only the contents of the web but also its political economy of visibility.

Organizations with the capacity to publish at scale can occupy more of the documentary environment. Search optimization, public relations, automated content generation, domain networks, and coordinated messaging can all expand a narrative's machine-readable presence. The danger is not limited to false statements. Selective truth, repeated framing, omission of context, and manufactured prominence can distort an informational environment without requiring every sentence to be false.

This is why the number of pages containing a claim cannot be treated as the number of independent sources supporting it.

A hundred websites may reproduce one press release. A thousand articles may be generated from one briefing document. Ten nominally separate organizations may share ownership, funding, staff, or political direction. Repetition measures presence. It does not establish independence, evidence, or truth.

A mature artificial-intelligence system must therefore reason not only over statements but also over relationships among statements. It must distinguish documents from sources, sources from evidence, and recurrence from corroboration. Research on claim-provenance graphs provides

one model for reconstructing how assertions originate and evolve, but this form of reasoning remains far more difficult than ordinary textual similarity.

The rise of synthetic publication makes the problem more urgent. Artificial-intelligence systems are increasingly producing material that becomes part of the public web. Later crawlers may collect this material, and later developers may incorporate it into new training corpora. The informational chain can become recursive: models learn from human writing, produce synthetic writing, and contribute that writing to the environment from which later models learn.

Research on recursive training demonstrates that uncontrolled reliance on model-generated data can progressively distort the learned distribution and suppress rare features. The danger is not necessarily immediate grammatical failure. A system may remain fluent while becoming narrower, more repetitive, and less faithful to the diversity of the original human record.

This gives historical Common Crawl releases an unexpected value. They may become reference layers preserving earlier stages of the web before synthetic text became pervasive. An archive collected for computational reuse may also become evidence of the changing boundary between human and machine expression.

The preservation of older web data does not automatically solve the problem. Earlier web corpora contain their own biases, misinformation, spam, and institutional inequalities. But they may preserve forms of human variation that later synthetic environments compress. The importance of such archives may therefore increase rather than decline as machine-generated publication expands.

Common Crawl also exposes an unresolved conflict between public access and computational reuse. Material placed on the open web is accessible to readers, but that does not settle every question about bulk archiving, redistribution, commercial training, retrieval, attribution, or synthetic imitation. Common Crawl itself states that material in its archive may remain subject to separate terms imposed by the owners of the captured content.

The current technical mechanisms for expressing permission are inadequate. A website operator may block CCBot through robots.txt, and Common Crawl states that its crawler checks those directives. Yet a binary crawling instruction cannot easily distinguish among search indexing, historical preservation, nonprofit research, commercial model training, retrieval-based quotation, or redistribution. It also cannot reliably remove copies already incorporated into archives and downstream datasets.

A more legitimate system will require purpose-specific permissions, dataset documentation, correction mechanisms, and clearer responsibility across the supply chain. It should be possible to know not merely whether a document was collected but how it was transformed, whether it

was retained, what relationship it has to other sources, and under what conditions it may be used.

Such accountability cannot be reduced to a universal prohibition against crawling. Common Crawl provides a genuine public benefit by making web-scale data available beyond the largest corporations. Its official account describes an archive exceeding 10 pebibytes, with approximately two billion or more pages in a typical crawl and new releases appearing roughly monthly. The June 2026 release alone contained 2.10 billion pages from 33.6 million registered domains.

Without shared archives, the ability to study the web would become more concentrated among firms capable of operating proprietary crawlers. Universities, nonprofit laboratories, journalists, public-interest researchers, and independent investigators would become increasingly dependent on corporate representations of the information environment.

The solution is therefore not to eliminate public data infrastructure but to make the surrounding system more accountable.

Common Crawl should be understood as analogous to a public observatory, but one directed toward a human-made and adversarial environment. Its observations should remain available for independent research. At the same time, the institutions using those observations to construct models should disclose their transformations more fully.

Dataset builders should document filtering, deduplication, language coverage, source concentration, synthetic-content detection, and known exclusions. Model developers should identify the major corpora used for training. Retrieval systems should cite evidence at the claim level and identify whether several cited pages derive from a common origin. Archives should connect obsolete pages with later corrections while preserving the historical record. Public institutions should possess enough computational capacity to audit web-scale datasets rather than relying solely on the assurances of their creators.

The objective cannot be a perfectly neutral corpus. Neutrality is impossible because every corpus embodies decisions about collection, representation, and exclusion. Nor can the objective be a model that contains only undisputed truth. Human knowledge includes hypotheses, testimony, interpretation, disagreement, fiction, error, satire, and historical change.

The more realistic objective is **provenance-complete plurality**.

A plural information system preserves competing voices rather than collapsing them prematurely into one consensus. A provenance-aware system records where those voices came from, whether they are independent, what evidence they possess, and how they changed over

time. An accountable system reveals the transformations through which documents became data and data became machine behavior.

This would allow artificial intelligence to say not only what appears to be the answer but also what kind of answer it is.

It could distinguish a settled measurement from a preliminary estimate, a primary source from commentary, an official assertion from independent verification, a widely repeated claim from a multiply confirmed finding, and a current fact from an obsolete one. It could preserve disagreement where the evidence remains divided rather than manufacturing certainty through fluent synthesis.

Common Crawl cannot produce this epistemic architecture by itself. It is an upstream archive, not a complete theory of knowledge. But its existence makes the need for such an architecture unmistakable. Once billions of pages can be collected, redistributed, filtered, and converted into model-training material, the governance of machine knowledge begins before model training.

It begins with publication.

It continues through crawl policy, archive structure, metadata preservation, dataset curation, source weighting, and correction.

By the time a user encounters an AI answer, the foundational decisions may already be several institutional layers and many years in the past.

The central lesson is therefore that artificial intelligence should not be judged only by the quality of its final outputs. It must also be judged by the integrity of its informational ancestry.

A fluent answer produced from opaque, distorted, or strategically manufactured data remains epistemically fragile. A powerful model built without provenance may amplify knowledge while weakening accountability. A system that retrieves current information without evaluating source relationships may accelerate access while making coordinated influence harder to recognize.

Common Crawl represents both the promise and the danger of open information at machine scale.

Its promise is that the public web can remain available as a shared object of study rather than becoming the exclusive property of dominant technology firms. It permits independent research, historical analysis, alternative dataset construction, and scrutiny of the informational foundations of artificial intelligence.

Its danger is that enormous scale can be mistaken for comprehensiveness, public availability for consent, repetition for consensus, textual cleanliness for epistemic quality, and statistical learning for knowledge.

The future of artificial intelligence will depend partly on whether societies preserve these distinctions.

If they do not, the public web may become an arena in which the actors capable of producing the most machine-readable language acquire disproportionate power over machine-mediated reality. Artificial intelligence will then reflect not the best-supported account of the world but the most successfully published, crawled, filtered, and repeated account.

If the distinctions are preserved, Common Crawl can become the foundation for a more open and inspectable intelligence ecosystem. It can support models that remain connected to documents, evidence, chronology, and human responsibility.

The decisive question is therefore not simply whether artificial intelligence learns from Common Crawl.

The decisive question is whether the path from the public web to machine knowledge remains visible enough to examine, contest, correct, and govern.

References

Common Crawl Foundation. “About.” The organization’s overview of its history, scale, technical infrastructure, data formats, and role in artificial-intelligence development.

Common Crawl Foundation. “Common Crawl Overview.” Description of the archive’s raw webpage data, metadata extracts, text extracts, indexes, and public hosting.

Common Crawl Foundation. “Frequently Asked Questions.” Documentation of crawl behavior, web sampling, robots.txt, sitemap support, and technical limitations.

Common Crawl Foundation. “June 2026 Crawl Archive Now Available.” Release statistics for crawl CC-MAIN-2026-25.

Common Crawl Foundation. “Terms of Use.” Statement concerning access to Common Crawl services and the separate rights or terms that may apply to captured material.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. “Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus.” 2021. The study analyzes the sources, filtering effects, machine-generated content, and demographic exclusions found in C4.

Penedo, Guilherme, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. “The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale.” 2024. The paper documents the construction of a 15-trillion-token corpus from 96 Common Crawl snapshots.

Seddik, Mohamed El Amine, Suei-Wen Chen, Soufiane Hayou, Pierre Youssef, and Merouane Debbah. “How Bad Is Training on Synthetic Data? A Statistical Analysis of Language Model Collapse.” 2024. The authors analyze the loss of distributional tails under recursive synthetic training and conditions under which mixtures of real and synthetic data may limit collapse.

Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. “AI Models Collapse When Trained on Recursively Generated Data.” *Nature*, 2024. The study describes degenerative effects arising when model-generated data contaminates the training sets of later models.

Zhang, Yi, Zachary Ives, and Dan Roth. “‘Who Said It, and Why?’ Provenance for Natural Language Claims.” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. The paper develops provenance graphs for tracing the origins and evolution of natural-language claims.